GLM multinomial hiérarchique

Jean Peyhardi [1,2], Catherine Trottier [1], Yann Guédon [2]

[1] *UM2, Institut de Mathématique et de Modélisation de Montpellier, 34095 Montpellier*
[2] *CIRAD, UMR AGAP et INRIA, Virtual Plants, F-34398 Montpellier*
jean.peyhardi@math.univ-montp2.fr

Résumé. L'objectif de ce travail est de proposer une écriture unifiée des GLMs (*Generalized Linear Model*) multinomiaux permettant d'aborder, en plus des cas classiques de données catégorielles nominales et ordinales, le cas des données catégorielles partiellement ordonnées. Cette écriture unifiée repose sur trois critères pour définir un GLM multinomial et permet de définir de nouveaux modèles. Le choix de la fonction de lien est aussi plus facilement interprétable dans ce nouveau cadre. Enfin cette écriture unifiée permet le développement d'une procédure d'estimation unique, basée sur l'algorithme des scores de Fisher. Sur cette base, nous proposons une modélisation hiérarchique à n étapes ($n \geq 2$), généralisant le *Nested Logit Model* de McFadden (1978), le *Two-Step Model* de Morawitz et Tutz (1990) et le *Partitioned Conditional Model* de Zhang et Ip (2012). L'idée est de bénéficier à chaque étape de la modélisation de cette écriture unifiée. Nous détaillons le cas $n = 2$. Il comprend une étape de partitionnement des catégories en sous-ensembles et une étape de conditionnement pour chaque sous-ensemble. Cette modélisation hiérarchique est illustré par le cas ce données partiellement ordonnées.

Mots-clés. Données catégorielles, GLM multinomial, Données partiellement ordonnées, Modèle hiérarchique.

Abstract. The objective of this work is to propose a unified writing of multinomial GLMs (Generalized linear Model) that enables to tackle not only the classical cases of nominal and ordinal categorical data but also the case of partially ordered categorical data. This unified writing relies on three criteria to define a multinomial GLM and enables to define new models. The choice of the link function is also more easily interpretable in this new setting. Finally, this unified writing enables to design a single estimation procedure based on Fisher scoring algorithm. On this basis, we propose a hierarchical modelling with n steps ($n \geq 2$) that generalizes the Nested Logit Model of McFadden (1978), the Two-Step Model of Morawitz and Tutz (1990) and the Partitioned Conditional Model of Zhang and Ip (2012). The idea is to benefit from this unified writing at each step. We detail the case $n = 2$. It involves a partitioning of categories in subsets and a conditioning step for each subset. This hierarchical modelling is illustrated with the case of partially ordered categorical data.

Keywords. Categorical data, Multinomial GLM, Partially ordered data, Hierarchical model.

1 Écriture unifiée des GLM multinomiaux

Soient Y la variable réponse, à J catégories ($J \geq 2$), et x le vecteur des variables explicatives. Par convention la dernière catégorie est choisie comme référence. La définition d'un GLM multinomial comprend une description de la fonction de lien g entre l'espérance $\pi = E[Y|X = x] = (\pi_1, ..., \pi_{J-1})^T$ et le prédicteur linéaire $\eta = (\eta_1, ..., \eta_{J-1})^T$. C'est un C^1-difféomorphisme de M dans S, où S est un ensemble ouvert de $\mathbb{R}^{J-1}$ et M est défini par:

$$M = \{\pi = (\pi_1, ..., \pi_{J-1}) \in]0,1[^{J-1} \ \mid \ \sum_{j=1}^{J-1} \pi_j < 1\}.$$

Nous proposons de décrire les différentes fonctions de lien $g = (g_1, \ldots, g_{J-1})$ usuelles, sous la forme suivante :

$$g_j = F^{-1} \circ r_j \ , \quad j = 1, \ldots, J-1,$$

où F est une fonction de répartion continue et strictement croissante sur $\mathbb{R}$ et $r = (r_1, \ldots, r_{J-1})^T$ est un C^1-difféomorphisme de M dans P, ensemble ouvert de $]0,1[^{J-1}$. Ainsi :

$$r_j(\pi) = F(\eta_j) \ , \quad j = 1, \ldots, J-1.$$

À partir des différents modèles existant dans la littérature (cf. Agresti (2002) et Fahrmeir et Tutz (2001)), nous décrivons les ratios ci-dessous et énumérons ensuite les lois et prédicteurs linéaires classiques :

Le ratio r

Reference

$$r_j(\pi) = \frac{\pi_j}{\pi_j + \pi_J} \ , \quad j = 1, \ldots, J-1,$$

Adjacent

$$r_j(\pi) = \frac{\pi_j}{\pi_j + \pi_{j+1}} \ , \quad j = 1, \ldots, J-1,$$

Sequential

$$r_j(\pi) = \frac{\pi_j}{\pi_j + \ldots + \pi_J} \ , \quad j = 1, \ldots, J-1,$$

Cumulative

$$r_j(\pi) = \pi_1 + \ldots + \pi_j \ , \quad j = 1, \ldots, J-1.$$

La loi de la variable latente F

Toutes les lois continues, avec pour support $\mathbb{R}$, conviennent. Les lois symétriques les plus souvent utilisées sont ***Logistic*** et ***Gaussian***. On peut ajouter la loi de ***Cauchy*** et de ***Student(d)*** (d étant le degré de liberté). Les lois asymétriques les plus souvent utilisées sont ***Gumbel max*** et ***Gumbel min***.

Le prédicteur linéaire η

Fahrmeir et Tutz (2001) écrivent η comme le produit d'une matrice de design Z et du vecteur de paramètres β. Les matrices de design classiques sont :

Intercept

$$Z = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix},$$

Shift

$$Z = \begin{pmatrix} 1 & & & x^T \\ & 1 & & x^T \\ & & \ddots & \vdots \\ & & 1 & x^T \end{pmatrix},$$

Threshold

$$Z = \begin{pmatrix} 1 & x^T & & & & \\ & & 1 & x^T & & \\ & & & & \ddots & \\ & & & & 1 & x^T \end{pmatrix},$$

Threshold-shift

$$Z = \begin{pmatrix} 1 & x_1^T & & & & x_2^T \\ & & 1 & x_1^T & & x_2^T \\ & & & & \ddots & \vdots \\ & & & & 1 \; x_1^T & x_2^T \end{pmatrix}.$$

Finalement, nous caractérisons un GLM multinomial particulier par le choix des trois critères (F, r, Z) et le résumons de la manière suivante :

$$r(\pi) = \mathbf{F}(Z\beta)$$

où $\mathbf{F}(\eta) = (F(\eta_1), \ldots, F(\eta_{J-1}))^T$.

La modélisation est ainsi simplifiée et plus facilement interprétable.

- Le choix de r correspond au choix d'un modèle totalement ordonné (*Sequential* ou *Cumulative*) ou non-ordonné (*Reference* ou *Adjacent*).
- Le choix de F modélise l évolution des probabilités $r_j(\pi)$ en fonction de x (**Fig. 1**). Une loi symétrique ou non, ou avec des queues plus ou moins lourdes permet de caractériser cette évolution.
- Le choix de Z correspond à aucun effet de x (*Intercept*), un effet commun sur toutes les catégories (*Shift*) ou un effet différent (*Threshold*).

Voici quelques exemples de modèles classiques (cf. Fahrmeir et Tutz, 2001) écrits sous cette forme :
Logit multinomial model :

$$P(Y = j) = \frac{\exp(\alpha_j + x^T\delta_j)}{1 + \sum_{k=1}^{J-1} \alpha_k + x^T\delta_k}, \; j = 1, \ldots, J-1 \quad \Leftrightarrow \quad (\textit{Reference, Logistic, Threshold})$$

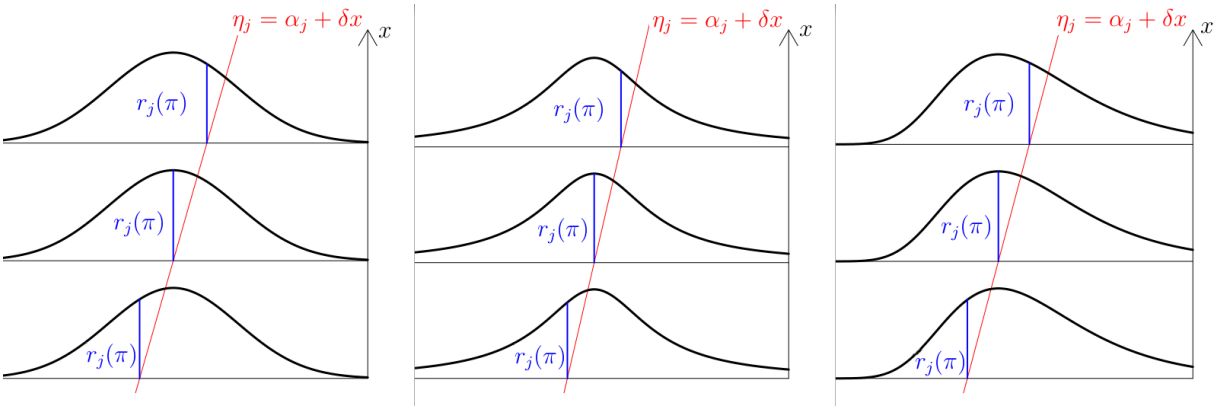


Fig. 1 : Répartion des probabilités $r_j(\pi)$ en fonction de $x \in \mathbb{R}$

a. Loi de Gauss **b.** Loi de Student(1) **c.** Loi de Gumbel max

Grouped Cox model :

$$P(Y \geq j) = \exp(-\exp(\alpha_j + x^T\delta)),\ j = 1, \ldots, J-1 \quad \Leftrightarrow \quad (\textit{Sequential, Gumbel min, Shift})$$

Odds proportional logit model :

$$\log\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) = \alpha_j + x^T\delta,\ j = 1, \ldots, J-1 \quad \Leftrightarrow \quad (\textit{Cumulative, Logistic, Shift})$$

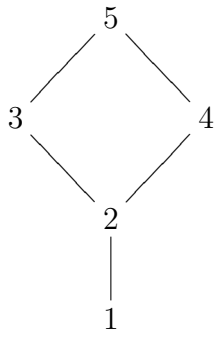
Nous obtenons aussi de nouveaux modèles grâce à certaines combinaisons non-explorées (comme (*Reference, Gumbel min, Shift*) par exemple) ou encore grâce à l'ajout des lois de *Cauchy* et de *Student(d)*.

Cette écriture unifiée permet l'implémentation d'une procédure d'estimation unique pour tous les GLMs multinomiaux. Elle est basée sur l'algorithme des scores de Fisher. Le calcul de la fonction score se décompose en une partie dépendante du triplet (r, F, Z) et une partie indépendante :

$$\frac{\partial l}{\partial \beta} = \underbrace{Z^T * \frac{\partial \mathbf{F}}{\partial \eta} * \frac{\partial \pi}{\partial r}}_{\text{dépendant}} * \underbrace{\mathrm{Cov}(Y)^{-1} * [y - \pi]}_{\text{indépendant}}.$$

Enfin cette écriture unifiée facilite la généralisation à des GLM multinomiaux hiérarchiques, qui permettent, par exemple, la prise en compte d'un ordre partiel sur les catégories.

2 GLM multinomial hiérarchique



Dans cette partie nous introduisons le GLM multinomial hiérarchique sur un exemple comprenant cinq catégories. Supposons qu'il existe un ordre partiel sur celles-ci, résumé par le treillis ci-contre. Zhang et Ip (2012) définissent des étapes de partitionnement des catégories en sous-ensembles, de telle sorte qu'à chaque étape les sous-ensembles soient totalement ordonnés (au sens faible) ou bien non-ordonnés. Ils proposent ainsi à chaque étape, respectivement le GLM ordinal (*Cumulative, Logistic, Shift*) ou bien le GLM nominal (*Reference, Logistic, Threshold*). Nous généralisons cette idée en proposant, à chaque étape, respectivement un GLM ordinal quelconque ou bien un GLM nominal quelconque.

Étape de partitionnement des catégories en sous-ensembles

Soit la partition suivante: $\mathcal{G}_1 = \{1\}$, $\mathcal{G}_2 = \{2\}$, $\mathcal{G}_3 = \{3, 4\}$, $\mathcal{G}_4 = \{5\}$. Notons p^* le vecteur des poids associé à cette partition:

$$p^* = (P(Y \in \mathcal{G}_1), P(Y \in \mathcal{G}_2), P(Y \in \mathcal{G}_3)) = (\pi_1, \pi_2, \pi_3 + \pi_4).$$

Les ensembles $\mathcal{G}_1$, $\mathcal{G}_2$, $\mathcal{G}_3$ et $\mathcal{G}_4$ étant totalement ordonnés (au sens fort), nous choisissons un GLM ordonné, c'est-à-dire un triplet (r^*, F^*, Z^*) avec un ratio $r^* = \textit{Sequential}$ ou $r^* = \textit{Cumulative}$.

Étape de conditionnement des catégories par rapport à chaque sous-ensemble

Les groupes $\mathcal{G}_1$, $\mathcal{G}_2$ et $\mathcal{G}_4$ ne contenant qu'une catégorie chacun, une deuxième étape de modélisation n'est pas nécessaire. Au contraire, pour le groupe $\mathcal{G}_3$, il reste une étape de différentiation. Notons p^3 le vecteur de probabilités de taille $|\mathcal{G}_3| - 1 = 1$:

$$p^3 = P(Y = 3 | Y \in \mathcal{G}_3) = \frac{\pi_3}{\pi_3 + \pi_4}.$$

Nous définissons alors un GLM pour cette loi conditionnelle par un triplet (r^3, F^3, Z^3).

Il existe des modèles hiérarchiques pour données catégorielles comme le *Nested Logit Model* de McFadden (1978) ou encore le *Two-Step Mode*l de Morawitz et Tutz (1990). Les deux permettent un changement du prédicteur linéaire à chaque étape. L'idée du *Nested Logit Model* est un peu différente : les variables explicatives peuvent changer d'un groupe à l'autre mais ne le doivent pas d'une catégorie à l'autre d'un même groupe. Cela revient à choisir des matrices de design dépendant de covariables différentes selon l'étape

de partitionnement ($Z^*(x^*)$) et selon l'étape de conditionnement par rapport à chacun des L groupes ($Z^{(1)}(x^{(1)}), \ldots, Z^{(L)}(x^{(L)})$). Pour une partition donnée, on peut résumer ces deux modèles comme suit :

$$\underline{\textit{Two-Step Model}} : \begin{cases} r^* = r^1 = \ldots = r^L = \textit{Cumulatif}, \\ F^* = F^1 = \ldots = F^L, \\ Z^*(x), Z^1(x), \ldots, Z^L(x). \end{cases}$$

$$\underline{\textit{Nested Logit Model}} : \begin{cases} r^* = r^1 = \ldots = r^L = \textit{Référence}, \\ F^* = F^1 = \ldots = F^L = \textit{Logistique}, \\ Z^*(x^*), Z^1(x^1), \ldots, Z^L(x^L). \end{cases}$$

On remarque que ces modèles hiérarchiques ne modifient pas r et F (c'est-à-dire la fonction de lien) d'une étape à l'autre. Au contraire, Zhang et Ip (2012) définissent le *Partitioned Conditional Model* qui modifie r entre *Reference* et *Cumulative*. On peut le résumer comme suit:

$$\underline{\textit{Partitioned Conditional Model}} : \begin{cases} r \in \{\textit{Reference, Cumulatif}\}^{L+1}, \\ F^* = F^1 = \ldots = F^L = \textit{Logistique}, \\ Z^*(x), Z^1(x), \ldots, Z^L(x). \end{cases}$$

Finalement Le GLM multinomial hiérarchique est modulable et la procédure d'estimation, basée sur l'algorithme des scores de Fisher, se décompose de la manière suivante:

$$\frac{\partial l}{\partial \beta} = \underbrace{Z^T * \frac{\partial \mathbf{F}}{\partial \eta} * \frac{\partial p}{\partial r} * \frac{\partial \pi}{\partial p}}_{\text{dépendant}} * \underbrace{\text{Cov}(Y)^{-1} * [y - \pi]}_{\text{indépendant}}$$

L'estimation des GLMs hiérarchiques pour données catégorielles, est ainsi simplifiée et leur comparaison se fait à l'aide de critères AIC et BIC par exemple. Nous illustrerons l'utilisation de ces modèles sur des données d'architecture de plantes.

Bibliographie

[1] Agresti, A. (2002), *Categorical data analysis*, John Wiley and Sons.
[2] Fahrmeir, L. et Tutz, G. (2001), *Multivariate statistical modelling based on generalized linear models*, Springer Verlag.
[3] McFadden, D. et al. (1978), *Modelling the choice of residential location*, Institute of Transportation Studies, University of California.
[4] Morawitz, B. et G. Tutz (1990), *Alternative parameterizations in business tendency surveys*, Mathematical Methods of Operations Research, Springer.
[5] Zhang, Q. et E. H. Ip (2012), *Generalized linear model for partially ordered data*, Statistics in Medicine.